Fast Solution

Protein Functional Sites Prediction (PFSP)

User Guide

Version 1.0

Developer & Author: Ruan Xiaoyang

Correspondence to: ruansun@163.com

ruansun@yahoo.cn

IMPORTANT NOTE:

PFSP DOES NOT support nucleotide sequence.

To reduce time to acceptable level in very time-consuming steps (such as estimation of pair wise Bayesian exact genetic distance for thousands of protein sequences), PFSP automatically FILTERS OUT ambiguous amino acid (AA) abbreviations including *X* (for unknown amino acid residue*)*, *B* (for aspartate or asparagines), *Z* (glutamate or glutamine), *O* (pyrrolysine), **-** (gap of indeterminate length), * (translation stop).

DO NOT separate PFSP main program from other shipped files in original package on first time use.

First Time Use

On first time use, the following files in PFSP package will be copied to C:\WINDOWS directory for next time use. PFSP will ask you to provide paths for these files if PFSP did not found them in the same folder.

Blocks.dat [Block file generate BLOSUM scoring matrices (if you accidentally lost it, search http://blocks.fhcrc.org/blocks/uploads/blosum/ for it)]

_ProFunSit_BLOSUM_RltvFreq [Pre-calculated BLOSUM relative frequency file with re-clustering standard from 100% to 14%]

_ProFunSit_t_s_thld [Pre-calculated BLAST hit score and extension score threshold]

PFSP_Userguide.pdf [User guide]

After first time use, you can copy/cut PFSP alone to other directory without above files

Log File

For each run of PFSP, a log file was saved to C:\ProFunSite.log. The log was automatically cleared the next time PFSP was initiated.

Table of contents

File Format

FASTA format. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. "|" was used follow ">" to separate gi-number, sequence source, accession number and locus information. PFSP do not have limitation on number of characters in each line. A newline character is necessary to separate description line from sequence data. The sequence ends if another ">"appears. An example sequence in FASTA format is:

>gi|532319|pir|TVFV2E|TVFV2E envelope protein

ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRTQIWQKHRTSNDSRGTNDPK-RIFFQRQWG**B**DPETANLWFNC

HGEFFYCKMDWFLN**Z**YLNNLTVDA**O**DHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKKTYAPPREGHLECTSTVTGMTVELNYIPKNRTNV

TLSPQIESIWAAELDRYKLVEITPIGFAPTEVRRYTGGHERQKRVPFV**XXXXXXXXXX**VQSQHLLAGILQQQKNLLAAVEAQQQMLKLTIWGVKNLLAAVE

*Note: In the above sequence, '-',**'X'**,**'B'**,**'Z'**,**'O'** and blank will be automatically filtered out.*

Custom format. This format enables users to use customized information as description line. A custom format sequence should also be started with ">" symbol. Other rules are same as that described in FASTA format. An example sequence in custom format is:

>input your description here and end with newline character

ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRTQIWQKHRTSNDS

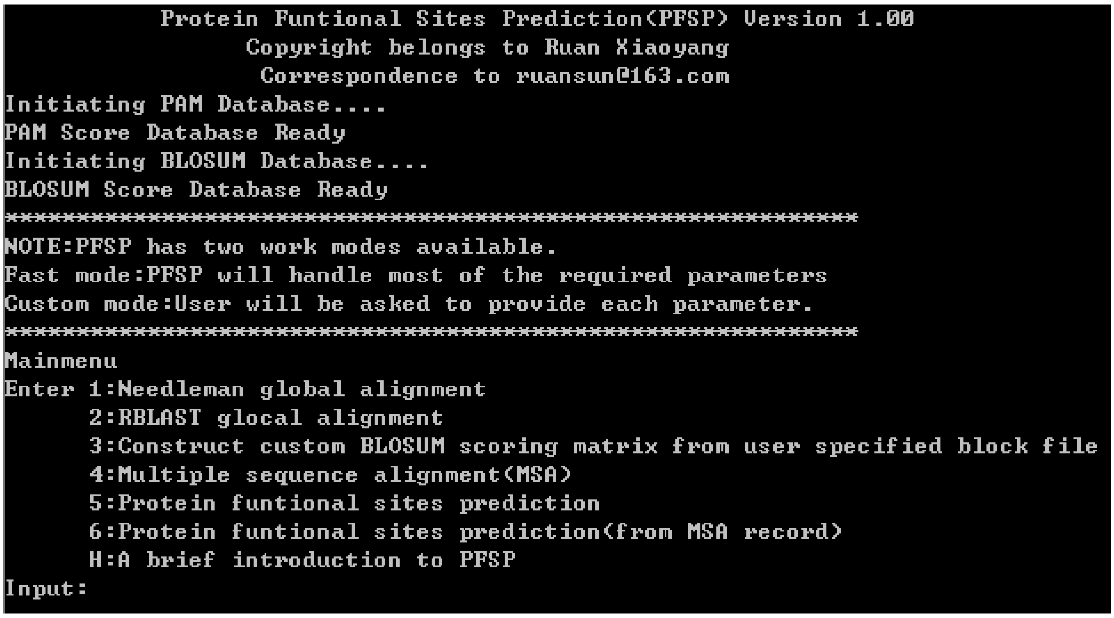
Note: DO NOT use mixed format in one file. This may cause serious problem.

Tips: Custom format is applicable to any file format start with ">" in their description line.

PFSP Options Overview

Initiating Scoring Database

Figure 1



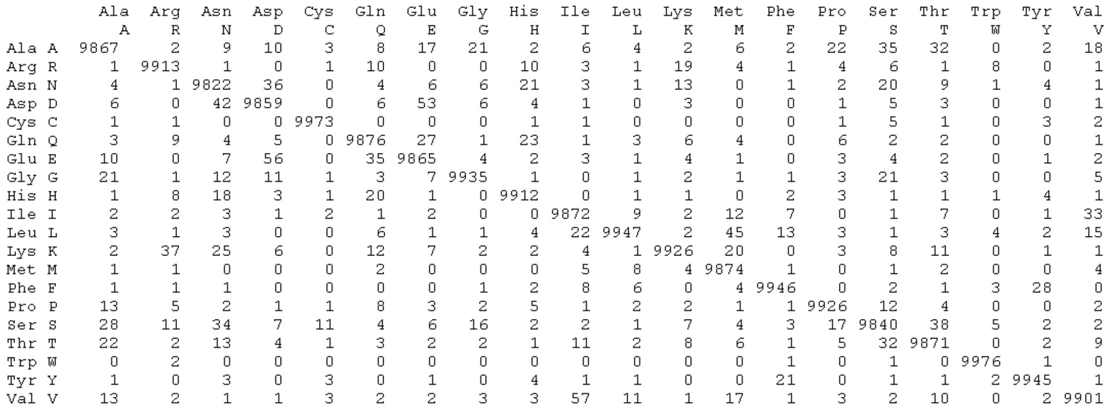
```
           Protein Funtional Sites Prediction(PFSP) Version 1.00
                 Copyright belongs to Ruan Xiaoyang
                  Correspondence to ruansun@163.com
Initiating PAM Database....
PAM Score Database Ready
Initiating BLOSUM Database....
BLOSUM Score Database Ready
************************************************************
NOTE:PFSP has two work modes available.
Fast mode:PFSP will handle most of the required parameters
Custom mode:User will be asked to provide each parameter.
************************************************************
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:
```

PFSP has two kinds of scoring matrices available. Dayhoff point accepted mutation (PAM) matrix [1] and blocks substitution matrix (BLOSUM) [2]. Both will be initiated at the startup of PFSP.

PAM matrices are based on global alignments of closely related proteins. The PAM 1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. Scores are derived from a mutation probability matrix where each element gives the probability of the amino acid in column X mutating to the amino acid in row Y after a particular evolutionary time, for example after 1 PAM, or 1% divergence. A PAM matrix is specific for a particular evolutionary distance, but may be used to generate matrices for greater evolutionary distances by multiplying it repeatedly by itself. However, at large evolutionary distances the information present in the matrix is essentially degenerated. It is rare that a PAM matrix would be used for an evolutionary distance any greater than 256 PAMs. Since all PAM scoring matrix at a longer evolutionary distance can be mathematically derived from PAM1, PFSP has a whole set of built-in scoring matrix from PAM1 to PAM400.

Whereas the PAM matrices have been developed from global alignments, the BLOSUM (BLOcks SUbstitution Matrix) matrices are based on local multiple alignments of more distantly related sequences. For instance, BLOSUM 62, the default matrix in BLAST, is a matrix calculated from comparisons of sequences with no less than 62% identity. Unlike PAM matrices, new BLOSUM matrices are never extrapolated from existing BLOSUM matrices, but are always based on local multiple alignments. So, the BLOSUM 80 matrix would be derived from a set of sequences having 80% sequence identity.

Figure 2



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

[top row shows original amino acid; left column shows replacement amino acid].

Mutation probability matrix for the evolutionary distance of 1 PAM (i.e. one accepted point mutation per 100 amino acids). An element of this matrix, [Mij], gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

BLOSUM scoring matrix was calculated from ***Blocks.dat*** file shipped with PFSP. However, PFSP does not bother to calculate BLOSUM scoring matrix every time. The pre-calculated relative frequency file ***_ProFunSit_BLOSUM_RltvFreq*** gives a fast solution to scoring matrices from BLOSUM100 to BLOSUM14. You will see later in this guide how to create your own BLOSUM scoring matrix.

Note: All scoring matrices in PFSP are scaled to 10×log[base 10]

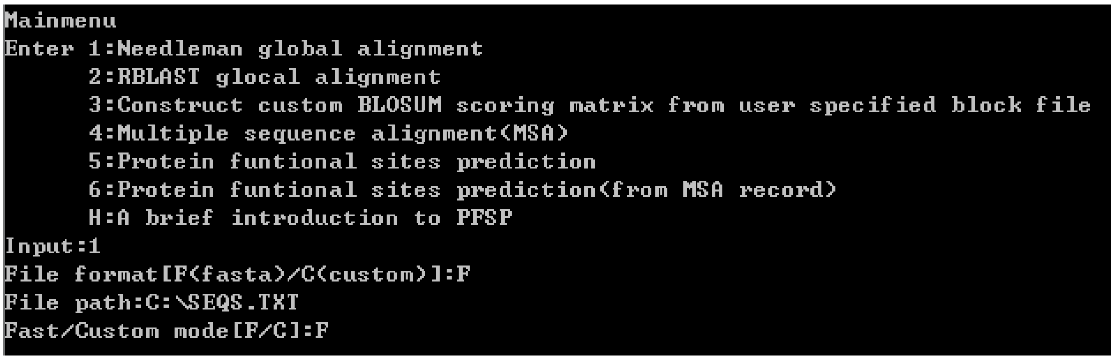
Work Modes

As you may notice in Figure 1, PFSP has two work modes available—Fast/Custom mode. Enter F (or f) for fast mode and C (of c) for custom mode (See Figure 3).

If you are not familiar with PFSP, or you have very little idea about how those different algorithms may affect the result, we suggest fast mode. See Table 1 for fast mode parameter list.

If you have carefully read this user guide and have a clear mind about which algorithm to use, we suggest custom mode. You will be better off if proper algorithms were chosen (See Table 2).

Figure 3



```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:1
File format[F(fasta)/C(custom)]:F
File path:C:\SEQS.TXT
Fast/Custom mode[F/C]:F
```

Example: choose fast mode in Needleman global alignment

Table 1 Algorithms used in fast mode

Option	Scoring database	Distance estimation type[b]	Distance estimation subtype	Gap penalty	Word length
Needleman global alignment	PAM[D][a]	Approximate[A]	Poisson[P]	-3	-
RBLAST glocal alignment	PAM[D]	Approximate[A]	Poisson[P]	-3	3
Multiple sequence alignment (MSA)	PAM[D]	Approximate[A]	Poisson[P]	-3	3
Protein functional sites prediction	PAM[D]	Approximate[A]	Poisson[P]	-3	3

[a] *Number inside square bracket is abbreviation*

[b] *Dayhoff PAM scoring rule was forced to be applied when evolutionary distance (PAM) was estimated. BLOSUM scoring rule was available only when using fixed scoring matrix*

Table 2 Available options in custom mode

Option	Scoring database	Distance estimation type [b]	Distance estimation subtype	Gap penalty	Word length	Bayesian step [f]	Gamma [g]	Seq number threshold
Needleman global alignment	PAM[D][a] BLOSUM[B]	Approx[A]	Simple[S][c]/Poisson[P]/Gamma[G]	Any [d]	-	-	Any	-
RBLAST glocal alignment	PAM[D] BLOSUM[B]	Approx[A] Bys Exact[E][e]	Simple[S]/Poisson[P]/Gamma[G] BLAST local[B]/Needleman global[N]	Any	3 or 4	10	Any	-
Multiple sequence alignment (MSA)	PAM[D]	Approx[A] Bys Exact[E]	Simple[S]/Poisson[P]/Gamma[G] BLAST local[B]/Needleman global[N]	Any	3 or 4	≤10	Any	-
Protein functional sites prediction	PAM[D]	Approx[A] Bys Exact[E]	Simple[S]/Poisson[P]/Gamma[G] BLAST local[B]/Needleman global[N]	Any	3 or 4	≤10	Any	≥10

[a] *Number inside square bracket is the abbreviation for that option*

[b] *Dayhoff PAM scoring rule was forced to be applied when evolutionary distance (PAM) was estimated. BLOSUM scoring rule was available only when using fixed scoring matrix*

[c] *Simple PAM distance=(1-(match ×2)/sum of length of two aligned sequences) ×100. Similarity% =(1-PAMdistance[Simple]/100) ×100%*

[d] *We suggest -3 for scoring matrix scaled to 10 ×log[base 10]*

[e] *Bayesian exact distance. Including BLAST local distance and Needleman global distance*

[f] *This term is used to adjust the balance between Bayesian speed and accuracy. Use 10 unless there is a good reason to use a lower value (lower speed also)*

[g] *When a = 2 is used, distance is close to Dayhoff's (1978) [1] PAM distance per site (0.01 PAM)*

Needleman Global Alignment

To have a better understanding of this section, the following knowledge is necessary.

PAM scoring matrix

BLOSUM scoring matrix

This algorithm was put forward by Needleman,S.B and Wunsch,C.D. [3] in 1970. This algorithm is a simple and beautiful (but less useful) way to find the maximum match between two sequences. You can try this algorithm if pair of sequences has very high similarity or if RBLAST glocal rule does not really meet your need (but we do not recommend it). This following example illustrated the steps needed to make pair wise Needleman global alignment in custom mode.

Sample sequences

>gi|159162458|pdb|1I6F|Insect-Specific Neurotoxin Variant 5 (Cse-V5)

KDGYPVDSKGCKLSCVANNYCDNQCKMKKASGGHCYAMSCYCEGLPENAKVSDSATNICG

>gi|158931147|sp|P60213.2|SC49A_TITCA Toxin Tc48b/Tc49a

KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTWSRSTNRCGR

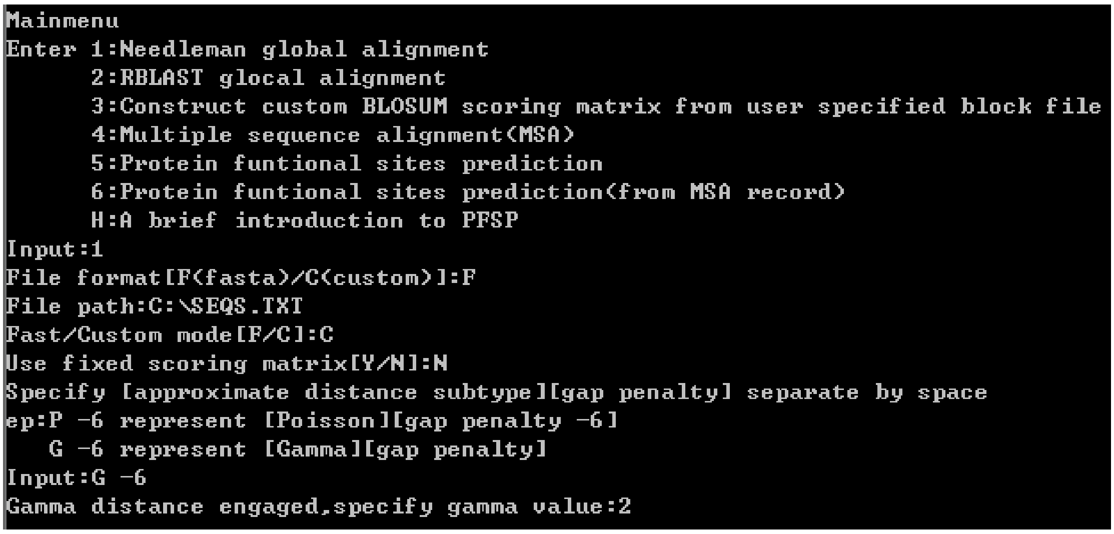
>gi|158705857|sp|P0C5F0.1|KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)

KIDGYPVDNWNCKRICWYNNKYCYDLCKGLKADSGYCWGWTLSCYCEGLPDNARIKRGGRCN

Store above Fasta format sequences in C:\seqs.txt

Figure 4

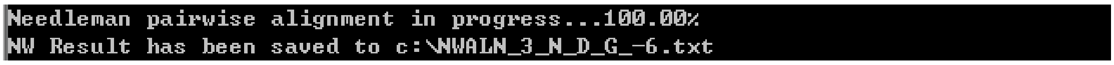
(Needleman global alignment; distance subtype: Gamma; gap penalty: -6; gamma value: 2)



```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:1
File format[F(fasta)/C(custom)]:F
File path:C:\SEQS.TXT
Fast/Custom mode[F/C]:C
Use fixed scoring matrix[Y/N]:N
Specify [approximate distance subtype][gap penalty] separate by space
ep:P -6 represent [Poisson][gap penalty -6]
   G -6 represent [Gamma][gap penalty]
Input:G -6
Gamma distance engaged,specify gamma value:2
```

Press enter and the following message appears

Figure 5

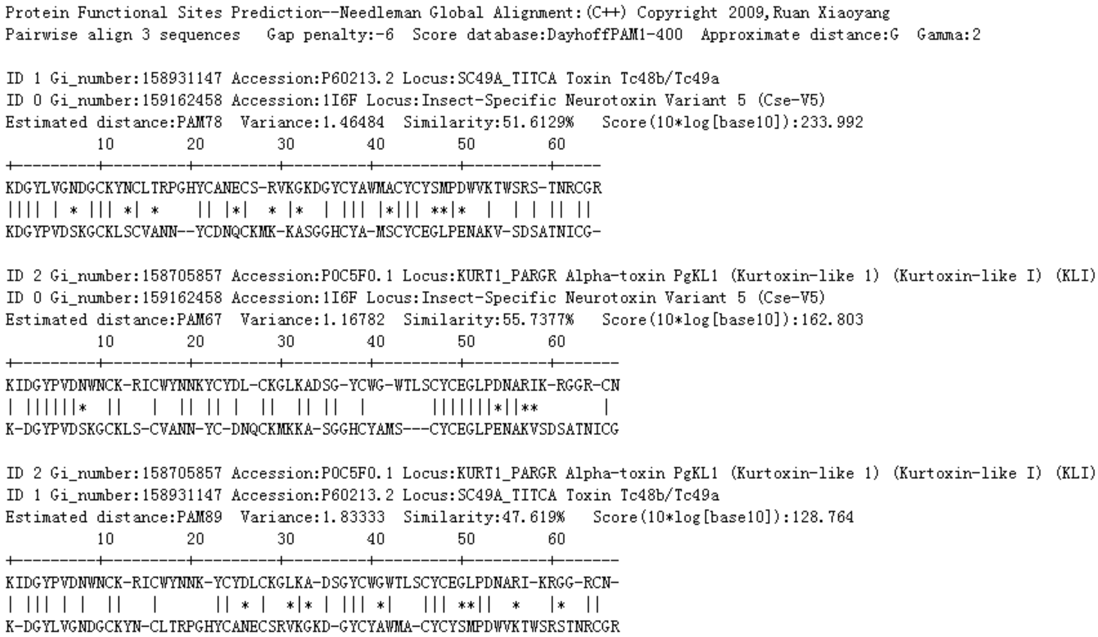


```
Needleman pairwise alignment in progress...100.00%
NW Result has been saved to c:\NWALN_3_N_D_G_-6.txt
```

C:\NWALN_3_N_D_G_-6.txt means [Needleman-Wunsch global alignment] _ [3 sequences] _ [scoring matrix not fixed] _ [Dayhoff PAM database] _ [Gamma distance] _ [gap penalty -6]

The alignment result will be automatically opened in notepad (Figure 6)

Figure 6



```
Protein Functional Sites Prediction--Needleman Global Alignment:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 3 sequences   Gap penalty:-6  Score database:DayhoffPAM1-400  Approximate distance:G  Gamma:2

ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM78  Variance:1.46484  Similarity:51.6129%   Score(10*log[base10]):233.992
         10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+-----
KDGYLVGNDGCKYNCLTRPGHYCANECS-RVKGKDGYCYAWMACYCYSMPDWVKTWSRS-TNRCGR
|||| | * ||| *| *    || |*|  * |*  | ||| |*||| **|*  |  | | || ||
KDGYPVDSKGCKLSCVANN--YCDNQCKMK-KASGGHCYA-MSCYCEGLPENAKV-SDSATNICG-

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM67  Variance:1.16782  Similarity:55.7377%   Score(10*log[base10]):162.803
         10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+-------
KIDGYPVDNWNCK-RICWYNNKYCYDL-CKGLKADSG-YCWG-WTLSCYCEGLPDNARIK-RGGR-CN
| ||||||*  ||   |  || || |  ||  || ||  |       |||||||*||**       |
K-DGYPVDSKGCKLS-CVANN-YC-DNQCKMKKA-SGGHCYAMS---CYCEGLPENAKVSDSATNICG

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
Estimated distance:PAM89  Variance:1.83333  Similarity:47.619%   Score(10*log[base10]):128.764
         10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+-------
KIDGYPVDNWNCK-RICWYNNK-YCYDLCKGLKA-DSGYCWGWTLSCYCEGLPDNARI-KRGG-RCN-
| ||| | |  ||   |      || * |  *|* | ||| *|   ||| **||  *   |*  ||
K-DGYLVGNDGCKYN-CLTRPGHYCANECSRVKGKD-GYCYAWMA-CYCYSMPDWVKTWSRSTNRCGR
```

The title shows the date and the alignment parameters. Sequence ID was defined as the order of appearance in the original file. "|"means identical, "" means the two aligned AA have score greater than 0 in corresponding scoring matrix. Gaps at the terminal of sequences were not scored.*

Since we have arbitrarily chosen gamma distance and arbitrarily set gamma parameter to 2, we are not sure if these parameters are optimal condition for the current sequences. Actually, several gaps with unreasonable length can be seen in the above alignment result. We suggest you to try fixed scoring matrix PAM250 or BLOSUM62. If none of these choices meet your requirement, use RBLAST instead.

Generally, Needleman rule was not preferred in that it maximizes the matching score on a global range, which may not truly represent the relationship between two sequences. The worst thing happens when the two sequences have very different lengths and one is a block of another. In such cases, Needleman rule forcibly elongates the shorter one by introducing in unreasonably long gaps. Although introducing in gap penalty term can to some extent compensate this apparent drawback, it does not provide a fundamental solution to common alignment works.

However, when sequence pairs are similar in length and have visually discernible similarity, Needleman global rule will prove similar result as other local/glocal algorithms. Example shown in RBLAST glocal alignment section best illustrated the difference between Needleman global rule and RBLAST glocal rule.

To allow for arbitrary length of gap (find the maximum number of match between two sequences) in an alignment, use gap penalty 0. Increase gap penalty results in decrease in gap length. When the absolute value of gap penalty equals to the maximum value in the scoring matrix, no gap was allowed.

For more information about Needleman global alignment

http://www2.cs.uh.edu/~zhenzhao/Review/alignment.htm

http://en.wikipedia.org/wiki/Needleman-Wunsch_algorithm

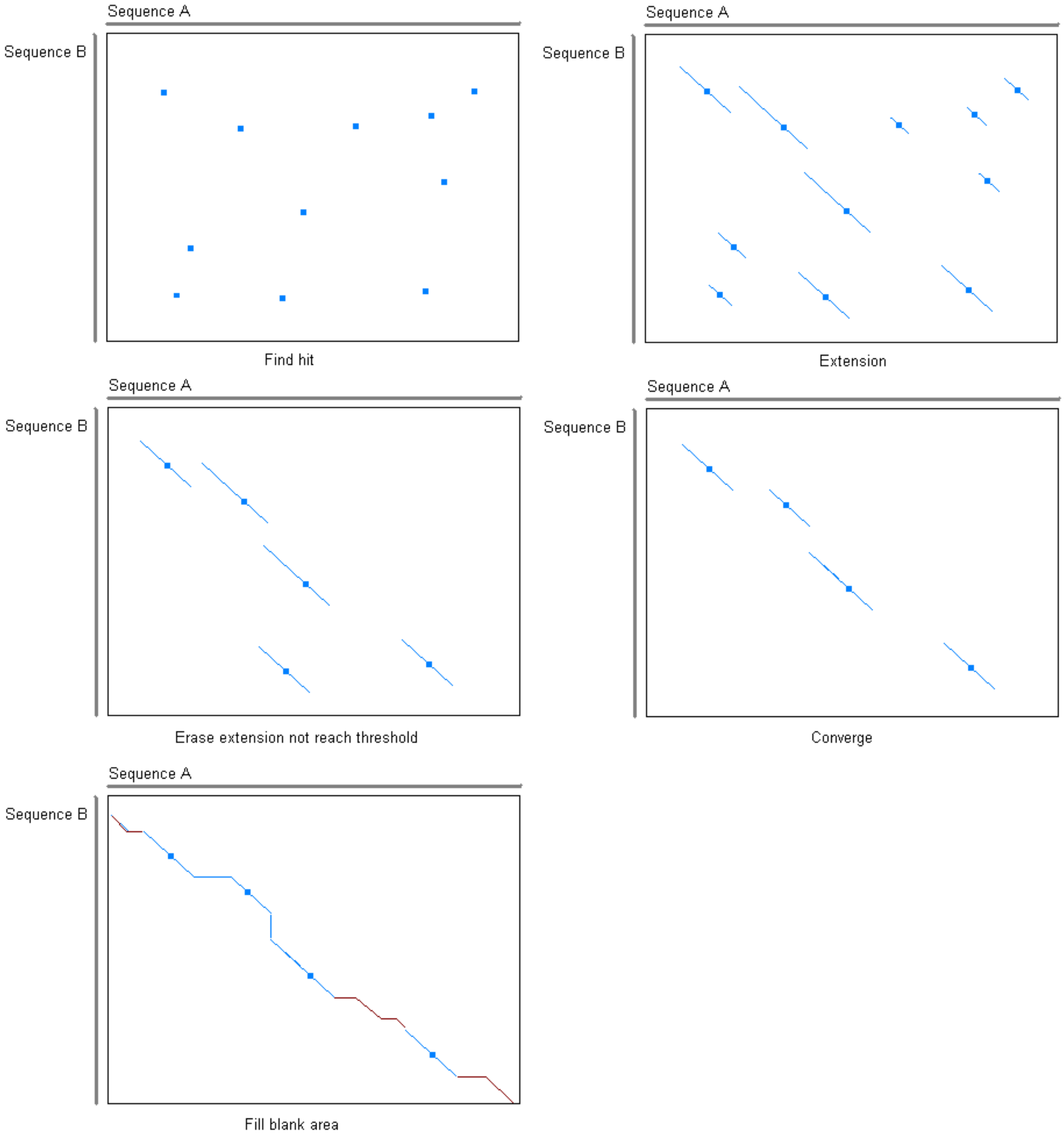
RBLAST Glocal Alignment

To have a better understanding of this section, the following knowledge is necessary.

Genetic distance [Approximate/Bayesian exact]

Basic local alignment searching tool (BLAST) algorithm was put forward by Altschul, S.F. et al [4] in 1990. It has now become one of the most widely used bioinformatics algorithm. The core idea of BLAST is to construct word libraries for both query and target (or database) sequences. Usually the word length is 3 or 4. The first appearance of a word adds a new record storing word and

Figure 7



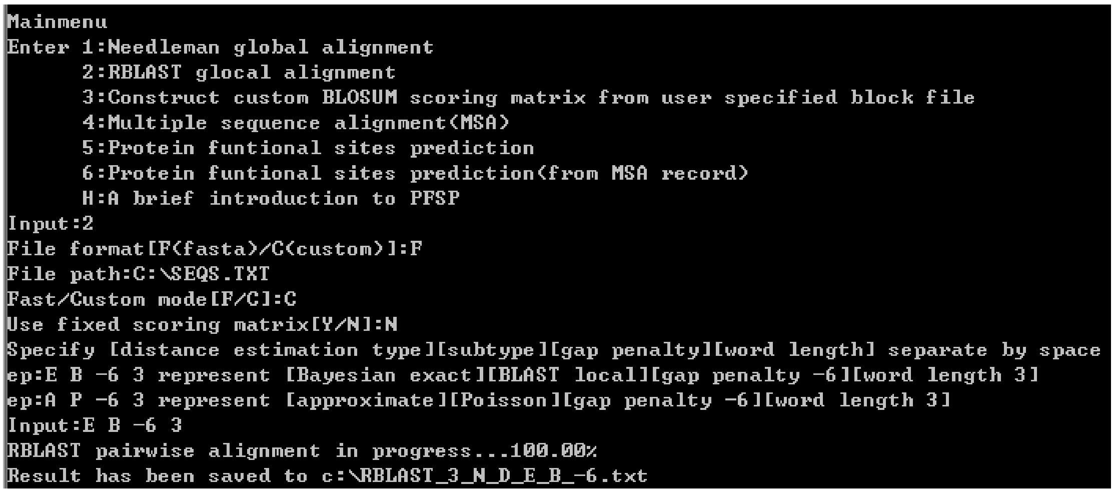
RBLAST glocal alignment work flow. Hit and extension score thresholds were pre-calculated and stored in _ProFunSit_t_s_thld. A new record will be added to this file if new alignment condition was encountered.

location information to the library. If same word record was found in the library, corresponding location repertoire was updated to contain the new location. Alignments of word pairs were then conducted to identify those pairs with matching score exceeding a certain threshold (called a hit). Hit was extended towards both ends until the score dropped below certain threshold. In PFSP, the threshold for hit was defined as the 95th percentile (a commonly used standard for statistical significance) of scores produced by N (N = 5000) randomly [a] generated word pairs of length W (word length) ([a] *Not really random, but depend on the natural frequency of AAs*). Different from regular BLAST, RBLAST does not control the extension by supervising the alignment score, it rather stops extension when W consecutive negative matching scores were encountered. An apparent benefit of this rule is that we have direct control of the maximum number of consecutive negative matches. So, for protein pairs that are distant from each other, a longer W may born better result than a shorter one. However, longer W may give very same result if protein pairs are evolutionarily close. The threshold used to disqualify extensions was defined as the 95th percentile of maximum extension scores produced by N (N = 5000) randomly generated protein sequence pairs of length S (S is the length of the shorter sequence in an alignment). So the hit score and extension score threshold vary with different word lengths and scoring matrices. After erasing disqualified extensions, the remaining segments were converged according to their locations and scores. Blank areas not covered in the BLAST step were filled by applying global alignment rule. The whole alignment process was described in figure 7.

The following example illustrated the use of RBLAST glocal alignment in custom mode

Figure 8

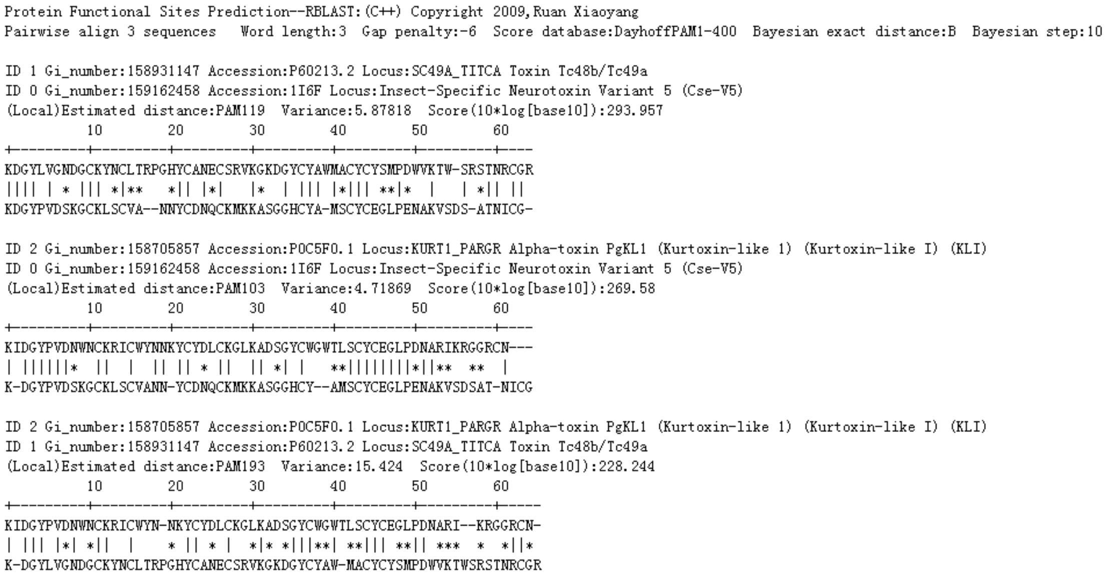
(RLAST global alignment; distance type: Bayesian exact; distance subtype: BLAST local; gap penalty: -6; word length: 3)



```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:2
File format[F(fasta)/C(custom)]:F
File path:C:\SEQS.TXT
Fast/Custom mode[F/C]:C
Use fixed scoring matrix[Y/N]:N
Specify [distance estimation type][subtype][gap penalty][word length] separate by space
ep:E B -6 3 represent [Bayesian exact][BLAST local][gap penalty -6][word length 3]
ep:A P -6 3 represent [approximate][Poisson][gap penalty -6][word length 3]
Input:E B -6 3
RBLAST pairwise alignment in progress...100.00%
Result has been saved to c:\RBLAST_3_N_D_E_B_-6.txt
```

C:\RBLAST_3_N_D_E_B_-6 means [RBLAST glocal alignment] _ [3 sequences] _ [scoring matrix not fixed] _ [Dayhoff PAM database] _ [Bayesian exact distance] _ [BLAST local distance] _ [gap penalty -6]

Figure 9



```
Protein Functional Sites Prediction--RBLAST:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 3 sequences   Word length:3  Gap penalty:-6  Score database:DayhoffPAM1-400  Bayesian exact distance:B  Bayesian step:10

ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
(Local)Estimated distance:PAM119  Variance:5.87818  Score(10*log[base10]):293.957
         10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+----
KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTW-SRSTNRCGR
|||| | * ||| *|**   *|| |*|   |*  | ||| |*||| **|*  |   | *|| ||
KDGYPVDSKGCKLSCVA--NNYCDNQCKMKKASGGHCYA-MSCYCEGLPENAKVSDS-ATNICG-

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
(Local)Estimated distance:PAM103  Variance:4.71869  Score(10*log[base10]):269.58
         10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+----
KIDGYPVDNWNCKRICWYNNKYCYDLCKGLKADSGYCWGWTLSCYCEGLPDNARIKRGGRCN---
| ||||||*  ||  |  || || * ||  || *| |   **||||||||*||**  **  |
K-DGYPVDSKGCKLSCVANN-YCDNQCKMKKASGGHCY--AMSCYCEGLPENAKVSDSAT-NICG

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
(Local)Estimated distance:PAM193  Variance:15.424  Score(10*log[base10]):228.244
         10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+-----
KIDGYPVDNWNCKRICWYN-NKYCYDLCKGLKADSGYCWGWTLSCYCEGLPDNARI--KRGGRCN-
| ||| |*| *||  |    * || * |  *|* *|||**| **||| **|| ***  *  *||*
K-DGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAW-MACYCYSMPDWVKTWSRSTNRCGR
```

Sequence data was shown in Needleman global alignment. Similarity is not available in Bayesian exact distance mode.

In this example, the result obtained by using RBLAST glocal rule is very similar to that by using Needle-man global rule. This is because all sequences in this example have similar length. Next we will illustrate the difference between RBLAST glocal rule and Needleman global rule. Consider the following sequences:

>father sequence

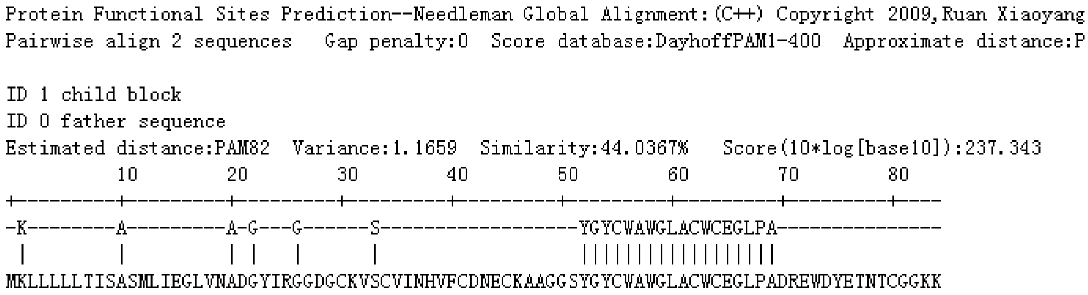
MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK

>child block

KAAGGSYGYCWAWGLACWCEGLPA

Child block is actually a segment intercepted from father sequence. Alignment of these two sequences by applying Needleman global rule (set gap penalty to 0 to allow for maximum match) generate the following result

Figure 10

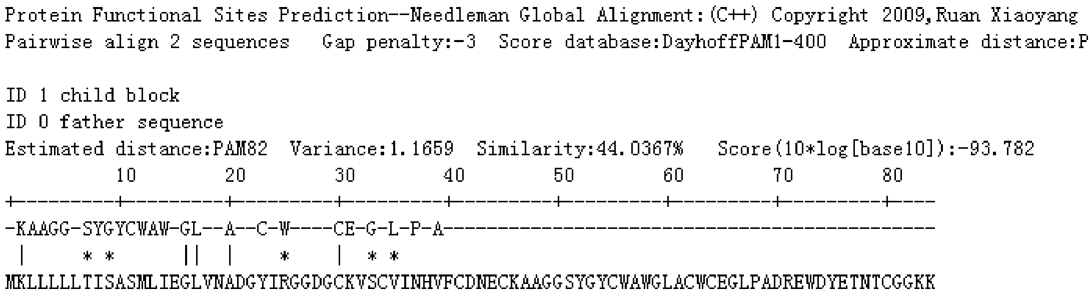


```
Protein Functional Sites Prediction--Needleman Global Alignment:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 2 sequences   Gap penalty:0  Score database:DayhoffPAM1-400  Approximate distance:P

ID 1 child block
ID 0 father sequence
Estimated distance:PAM82  Variance:1.1659  Similarity:44.0367%   Score(10*log[base10]):237.343
         10        20        30        40        50        60        70        80
+---------+---------+---------+---------+---------+---------+---------+---------+----
-K--------A---------A-G---G------S------------------YGYCWAWGLACWCEGLPA---------------
 |        |         | |   |      |                  ||||||||||||||||||
MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK
```

This result obviously is not what we want. You may argue that set gap penalty to -3 or -6 will get rid of those unreasonably long gap. Let's try -3

Figure 11

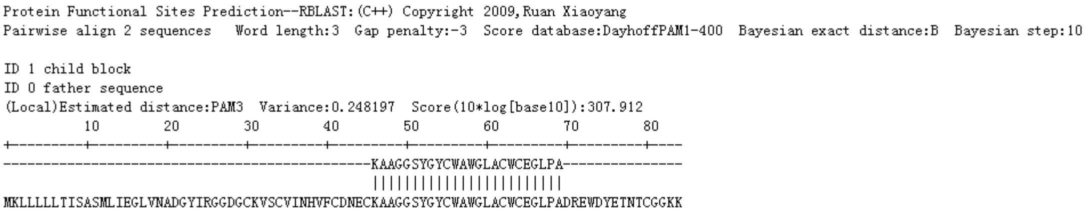


```
Protein Functional Sites Prediction--Needleman Global Alignment:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 2 sequences   Gap penalty:-3  Score database:DayhoffPAM1-400  Approximate distance:P

ID 1 child block
ID 0 father sequence
Estimated distance:PAM82  Variance:1.1659  Similarity:44.0367%   Score(10*log[base10]):-93.782
         10        20        30        40        50        60        70        80
+--------+---------+---------+---------+---------+---------+---------+---------+----
-KAAGG-SYGYCWAW-GL--A--C-W----CE-G-L-P-A---------------------------------------------
 |     * *      ||  |    *    |  * *
MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK
```

This one is even worse. Actually, all AAs of the child block sequence will be pushed into a pile at the left end of the alignment if gap penalty exceeded -11. However, RBLAST glocal rule will do the job excellently

Figure 12



```
Protein Functional Sites Prediction--RBLAST:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 2 sequences   Word length:3  Gap penalty:-3  Score database:DayhoffPAM1-400  Bayesian exact distance:B  Bayesian step:10

ID 1 child block
ID 0 father sequence
(Local)Estimated distance:PAM3  Variance:0.248197  Score(10*log[base10]):307.912
         10        20        30        40        50        60        70        80
+--------+---------+---------+---------+---------+---------+---------+---------+----
----------------------------------------------KAAGGSYGYCWAWGLACWCEGLPA---------------
                                              ||||||||||||||||||||||||
MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK
```

In PFSP, RBLAST is an important component of multiple sequence alignment (MSA). The global rule used in RBLAST to fill the blank areas not covered by local alignment is slightly different from the Needleman global alignment introduced above. The main difference is that RBLAST always ensure maximum match when invoke global rule to align blank areas. The gap penalty term does not control the gap length. This rule has both advantages and disadvantages. The advantage is that it helps to increase the number of identical AAs in MSA result, which might be helpful to identify functionally important conservative sites. The disadvantage is that gaps with unreasonable length sometimes deviate from the true situation in the natural evolutionary process.

For more information about BLAST and glocal alignment

http://blast.ncbi.nlm.nih.gov/Blast.cgi

http://en.wikipedia.org/wiki/BLAST

Multiple Sequence Alignment (MSA)

To have a better understanding of this section, the following knowledge is necessary.

RBLAST glocal alignment

Genetic distance

Phylogenetic tree

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In general, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in figure 11 illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex to produce. Most multiple sequence alignment programs use heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive.

In PFSP, MSA is a vital step to protein function sites prediction. As such the quality of MSA result determined the reliability of functional sites predicted. PFSP has incorporated a panel of algorithms that enable you to balance between accuracy and speed (Table 2). To obtain the fastest speed, use approximate distance->simple distance. To obtain the highest accuracy, use Bayesian exact distance->BLAST local distance. The work flow of MSA is as follow.

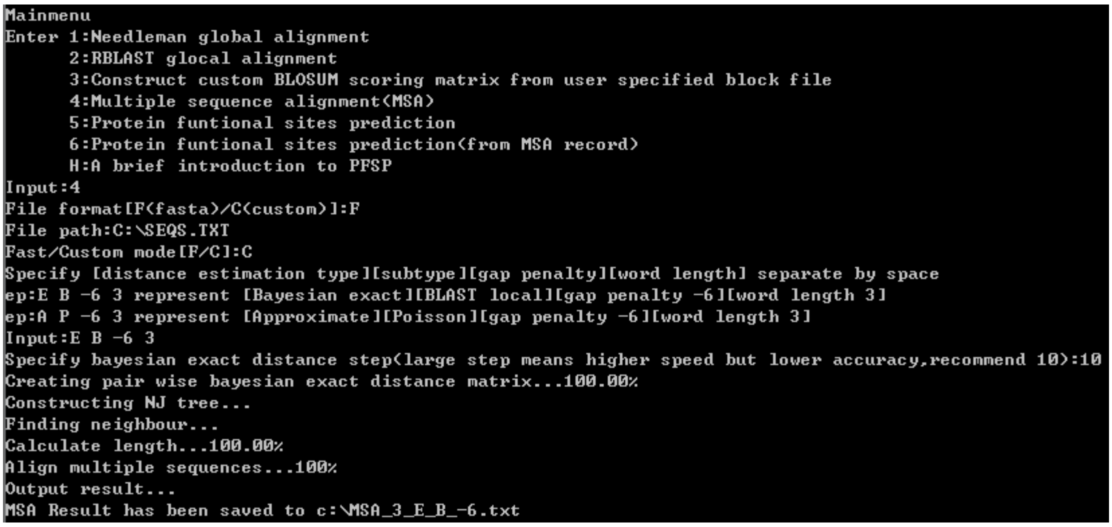
Import File -> Pair-wise Evolutionary Distance Estimation -> Phylogenetic tree construction -> Sequence alignment

The most time-consuming step is pair-wise evolutionary distance estimation, which has a time complexity of $O\ (n^2)$. The following examples illustrated multiple align 3 protein sequences and 270 protein sequences in custom mode (Sample file is available in the package

SAMPLE_SEQS.txt).

Figure 13

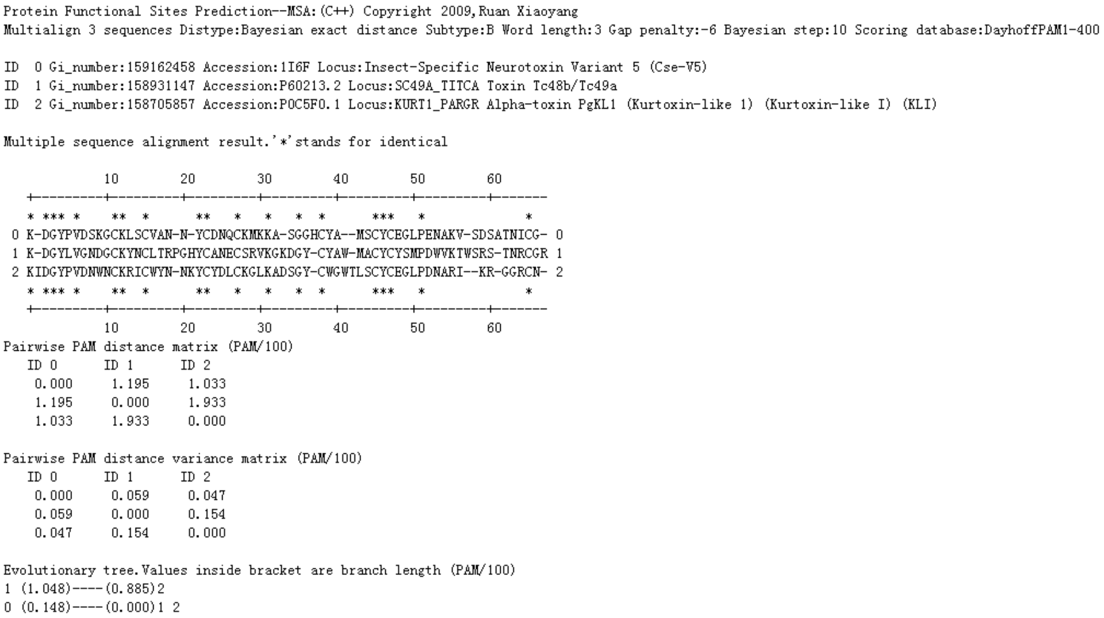
(MSA; distance type: Bayesian exact; distance subtype: BLAST local; gap penalty: -6; word length: 3)



```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:4
File format[F(fasta)/C(custom)]:F
File path:C:\SEQS.TXT
Fast/Custom mode[F/C]:C
Specify [distance estimation type][subtype][gap penalty][word length] separate by space
ep:E B -6 3 represent [Bayesian exact][BLAST local][gap penalty -6][word length 3]
ep:A P -6 3 represent [Approximate][Poisson][gap penalty -6][word length 3]
Input:E B -6 3
Specify bayesian exact distance step(large step means higher speed but lower accuracy,recommend 10):10
Creating pair wise bayesian exact distance matrix...100.00%
Constructing NJ tree...
Finding neighbour...
Calculate length...100.00%
Align multiple sequences...100%
Output result...
MSA Result has been saved to c:\MSA_3_E_B_-6.txt
```

C:\MSA_3_E_B_-6 means [MSA] _ [3 sequences] _ [Bayesian exact distance] _ [BLAST local distance] _ [gap penalty -6]

Figure 14



```
Protein Functional Sites Prediction--MSA:(C++) Copyright 2009,Ruan Xiaoyang
Multialign 3 sequences Distype:Bayesian exact distance Subtype:B Word length:3 Gap penalty:-6 Bayesian step:10 Scoring database:DayhoffPAM1-400

ID  0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
ID  1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
ID  2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)

Multiple sequence alignment result.'*'stands for identical

            10        20        30        40        50        60
   +---------+---------+---------+---------+---------+---------+-------
   * *** *    **  *      **   *   *   *  *      ***   *             *
 0 K-DGYPVDSKGCKLSCVAN-N-YCDNQCKMKKA-SGGHCYA--MSCYCEGLPENAKV-SDSATNICG- 0
 1 K-DGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGY-CYAW-MACYCYSMPDWVKTWSRS-TNRCGR 1
 2 KIDGYPVDNWNCKRICWYN-NKYCYDLCKGLKADSGY-CWGWTLSCYCEGLPDNARI--KR-GGRCN- 2
   * *** *    **  *      **   *   *   *  *      ***   *             *
   +---------+---------+---------+---------+---------+---------+-------
            10        20        30        40        50        60
Pairwise PAM distance matrix (PAM/100)
   ID 0      ID 1      ID 2
    0.000     1.195     1.033
    1.195     0.000     1.933
    1.033     1.933     0.000

Pairwise PAM distance variance matrix (PAM/100)
   ID 0      ID 1      ID 2
    0.000     0.059     0.047
    0.059     0.000     0.154
    0.047     0.154     0.000

Evolutionary tree.Values inside bracket are branch length (PAM/100)
1 (1.048)----(0.885)2
0 (0.148)----(0.000)1 2
```

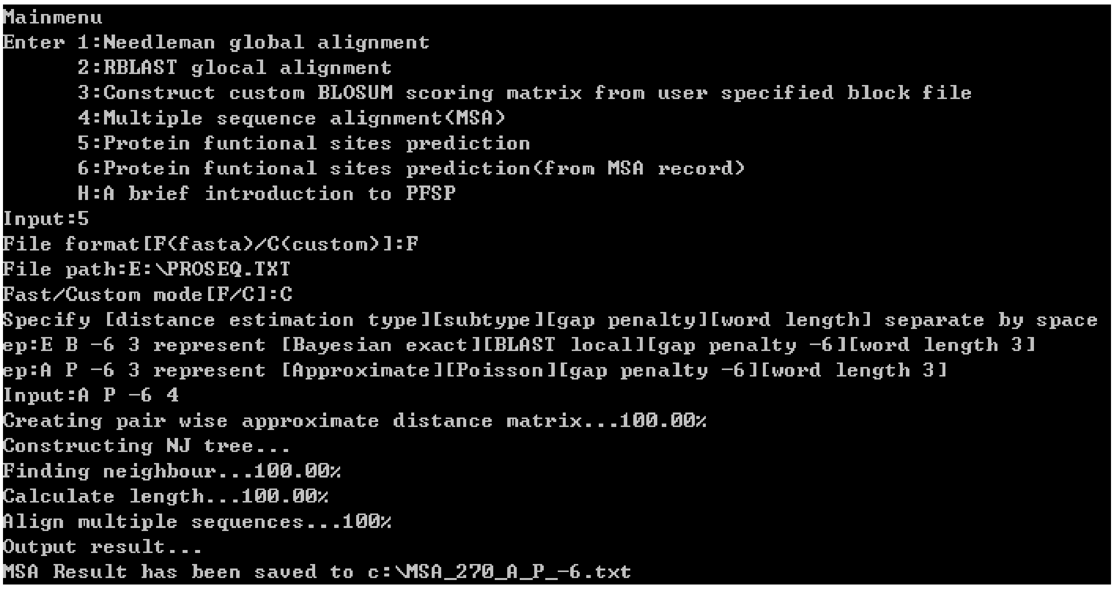
Multiple align 3 sequences by using Bayesian exact distance. The output result includes PAM distance matrix, distance variance matrix and evolutionary tree. Neighbors were connected by "----". A MSA record file named MSA_3_E_B_-6_record was automatically saved to c:\ directory. The record file enables you to do functionally sites prediction directly from option 6 rather than have to align the sequences again (Note: at least 20 sequences are required to enable functional sites prediction).

Figure 15 and 16 illustrated multiple alignment of 270 sequences in Fasta format. To save time, we chosen Poisson algorithm to estimate approximate distance. Word length was set to 4.

Generally, when sequence number is large, different distance estimation algorithms may generate visually discernable different alignment results. As we can see in figure 12, there are 2 positions have identical AA (marked by "*"). However, if we apply Bayesian exact distance, actually no identical AA will be identified. Does this mean Poisson approximate algorithm performs better than Bayesian exact algorithm? The answer is NO. The quality of a MSA primarily depends on the number of conservative site found. In this example, as you will noticed in Figure 15

Figure 15

(MSA; distance type: approximate; distance subtype: Poisson; gap penalty: -6; word length: 4)



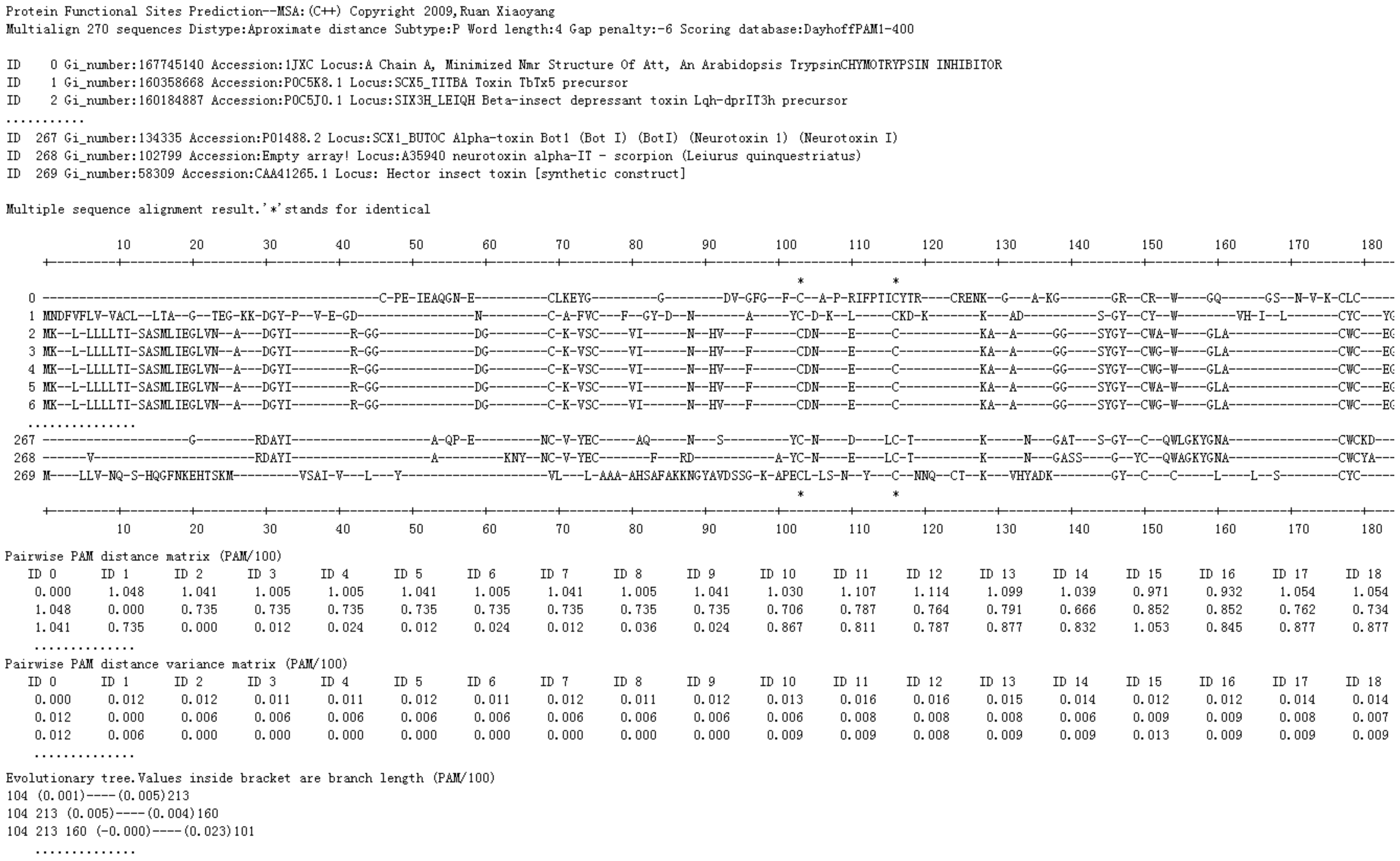
```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:5
File format[F(fasta)/C(custom)]:F
File path:E:\PROSEQ.TXT
Fast/Custom mode[F/C]:C
Specify [distance estimation type][subtype][gap penalty][word length] separate by space
ep:E B -6 3 represent [Bayesian exact][BLAST local][gap penalty -6][word length 3]
ep:A P -6 3 represent [Approximate][Poisson][gap penalty -6][word length 3]
Input:A P -6 4
Creating pair wise approximate distance matrix...100.00%
Constructing NJ tree...
Finding neighbour...100.00%
Calculate length...100.00%
Align multiple sequences...100%
Output result...
MSA Result has been saved to c:\MSA_270_A_P_-6.txt
```

C:\MSA_270_A_P_-6 means [MSA] _ [270 sequences] _ [approximate distance] _ [Poisson distance] _ [gap penalty -6]

Poisson algorithm identified 13 sites with conservative score exceeding 95th percentile and 12 sites exceeding 90th. However, Bayesian algorithm will identify 14 (95th) and 12 (90th) sites. This difference is not statistically significant, but it reminds us that the number of "*" (identical AA) do not reflect the quality of MSA.

For each run of MSA, a record file containing the MSA result was saved to C:\ directory. This file can be imported by PFSP for further analysis (option 6). The record file was preceded by a line specifies the row and column number information. Then follows identical AA information marked by "*".

Figure 16



```
Protein Functional Sites Prediction--MSA:(C++) Copyright 2009,Ruan Xiaoyang
Multialign 270 sequences Distype:Aproximate distance Subtype:P Word length:4 Gap penalty:-6 Scoring database:DayhoffPAM1-400

ID    0 Gi_number:167745140 Accession:1JXC Locus:A Chain A, Minimized Nmr Structure Of Att, An Arabidopsis TrypsinCHYMOTRYPSIN INHIBITOR
ID    1 Gi_number:160358668 Accession:P0C5K8.1 Locus:SCX5_TITBA Toxin TbTx5 precursor
ID    2 Gi_number:160184887 Accession:P0C5J0.1 Locus:SIX3H_LEIQH Beta-insect depressant toxin Lqh-dprIT3h precursor
...........
ID  267 Gi_number:134335 Accession:P01488.2 Locus:SCX1_BUTOC Alpha-toxin Bot1 (Bot I) (BotI) (Neurotoxin 1) (Neurotoxin I)
ID  268 Gi_number:102799 Accession:Empty array! Locus:A35940 neurotoxin alpha-IT - scorpion (Leiurus quinquestriatus)
ID  269 Gi_number:58309 Accession:CAA41265.1 Locus: Hector insect toxin [synthetic construct]

Multiple sequence alignment result.'*'stands for identical

              10        20        30        40        50        60        70        80        90       100       110       120       130       140       150       160       170       180
    +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+----
                                                                                                           *            *
  0 ----------------------------------------------C-PE-IEAQGN-E----------CLKEYG---------G--------DV-GFG--F-C--A-P-RIFPTICYTR----CRENK--G---A-KG-------GR--CR--W----GQ------GS--N-V-K-CLC-----
  1 MNDFVFLV-VACL--LTA--G--TEG-KK-DGY-P--V-E-GD----------------N---------C-A-FVC---F--GY-D--N-------A-----YC-D-K--L-----CKD-K-------K---AD----------S-GY--CY--W--------VH-I--L-------CYC---YG
  2 MK--L-LLLLTI-SASMLIEGLVN--A---DGYI--------R-GG-------------DG--------C-K-VSC----VI------N--HV---F------CDN----E-----C-----------KA--A-----GG----SYGY--CWA-W----GLA---------------CWC---EG
  3 MK--L-LLLLTI-SASMLIEGLVN--A---DGYI--------R-GG-------------DG--------C-K-VSC----VI------N--HV---F------CDN----E-----C-----------KA--A-----GG----SYGY--CWG-W----GLA---------------CWC---EG
  4 MK--L-LLLLTI-SASMLIEGLVN--A---DGYI--------R-GG-------------DG--------C-K-VSC----VI------N--HV---F------CDN----E-----C-----------KA--A-----GG----SYGY--CWG-W----GLA---------------CWC---EG
  5 MK--L-LLLLTI-SASMLIEGLVN--A---DGYI--------R-GG-------------DG--------C-K-VSC----VI------N--HV---F------CDN----E-----C-----------KA--A-----GG----SYGY--CWA-W----GLA---------------CWC---EG
  6 MK--L-LLLLTI-SASMLIEGLVN--A---DGYI--------R-GG-------------DG--------C-K-VSC----VI------N--HV---F------CDN----E-----C-----------KA--A-----GG----SYGY--CWG-W----GLA---------------CWC---EG
  ...............
267 --------------------G--------RDAYI-------------------A-QP-E---------NC-V-YEC-----AQ-----N---S---------YC-N----D----LC-T---------K-----N---GAT---S-GY--C--QWLGKYGNA---------------CWCKD---
268 ------V----------------------RDAYI-------------------A---------KNY--NC-V-YEC-------F---RD-----------A-YC-N----E----LC-T---------K-----N---GASS----G--YC--QWAGKYGNA---------------CWCYA---
269 M----LLV-NQ-S-HQGFNKEHTSKM---------VSAI-V---L---Y--------------------VL---L-AAA-AHSAFAKKNGYAVDSSG-K-APECL-LS-N--Y---C--NNQ--CT--K---VHYADK--------GY--C---C-----L----L--S--------CYC-----
                                                                                                           *            *
    +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+----
              10        20        30        40        50        60        70        80        90       100       110       120       130       140       150       160       170       180
Pairwise PAM distance matrix (PAM/100)
   ID 0      ID 1      ID 2      ID 3      ID 4      ID 5      ID 6      ID 7      ID 8      ID 9      ID 10     ID 11     ID 12     ID 13     ID 14     ID 15     ID 16     ID 17     ID 18
    0.000     1.048     1.041     1.005     1.005     1.041     1.005     1.041     1.005     1.041     1.030     1.107     1.114     1.099     1.039     0.971     0.932     1.054     1.054
    1.048     0.000     0.735     0.735     0.735     0.735     0.735     0.735     0.735     0.735     0.706     0.787     0.764     0.791     0.666     0.852     0.852     0.762     0.734
    1.041     0.735     0.000     0.012     0.024     0.012     0.024     0.012     0.036     0.024     0.867     0.811     0.787     0.877     0.832     1.053     0.845     0.877     0.877
    ..............
Pairwise PAM distance variance matrix (PAM/100)
   ID 0      ID 1      ID 2      ID 3      ID 4      ID 5      ID 6      ID 7      ID 8      ID 9      ID 10     ID 11     ID 12     ID 13     ID 14     ID 15     ID 16     ID 17     ID 18
    0.000     0.012     0.012     0.011     0.011     0.012     0.011     0.012     0.011     0.012     0.013     0.016     0.016     0.015     0.014     0.012     0.012     0.014     0.014
    0.012     0.000     0.006     0.006     0.006     0.006     0.006     0.006     0.006     0.006     0.006     0.008     0.008     0.008     0.006     0.009     0.009     0.008     0.007
    0.012     0.006     0.000     0.000     0.000     0.000     0.000     0.000     0.000     0.000     0.009     0.009     0.008     0.009     0.009     0.013     0.009     0.009     0.009
    ..............
Evolutionary tree.Values inside bracket are branch length (PAM/100)
104 (0.001)----(0.005)213
104 213 (0.005)----(0.004)160
104 213 160 (-0.000)----(0.023)101
    ..............
```

Part of multiple sequences alignment result. Incomplete sequence source information was labeled by “Empty array!

Protein Functional Sites Prediction

To have a better understanding of this section, the following knowledge is necessary.

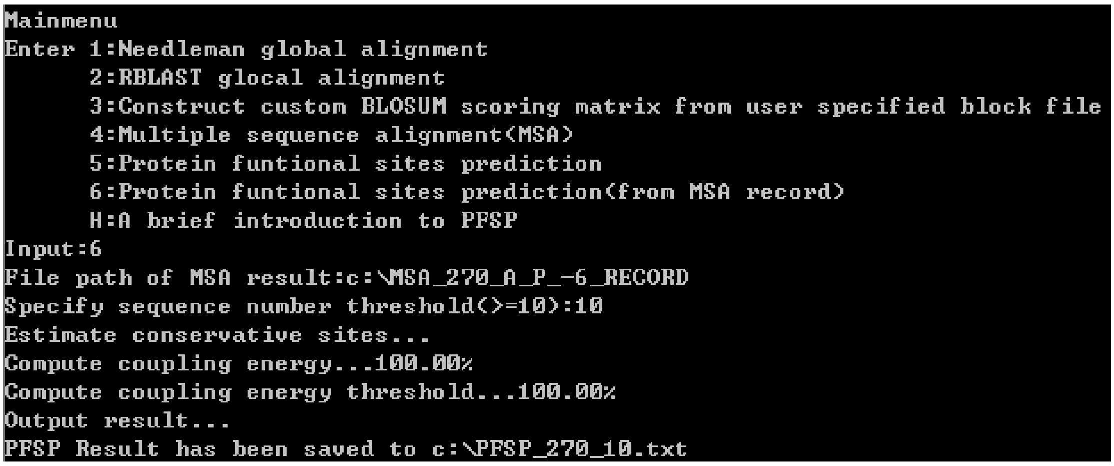
Multiple sequence alignment

The realization of functional sites prediction primarily depends on the coupling energy fluctuation bring about by mutation at specific location. The idea of statistical coupling energy has been systematical described by Steve W. Lockless, *et al*.[5]. PFSP added to their analyses an average coupling energy calculated by averaging over the coupling energy changes at the rest of sites when one site mutated. PFSP also output the estimated coupling energy for every site in the most intuitive way. These results could be used as aids to identify potential functional sites and help biologists to optimize their mutation plan (as long as enough sequence is available to justify the statistics).

Let's first see an example. There are two ways to do functional sites prediction in PFSP. One is from raw sequences and another is from MSA record. In the following example, we start from MSA record generated from previous section.

Figure 17

(PFSP from MSA record; sequence number threshold: 10)



```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:6
File path of MSA result:c:\MSA_270_A_P_-6_RECORD
Specify sequence number threshold(>=10):10
Estimate conservative sites...
Compute coupling energy...100.00%
Compute coupling energy threshold...100.00%
Output result...
PFSP Result has been saved to c:\PFSP_270_10.txt
```

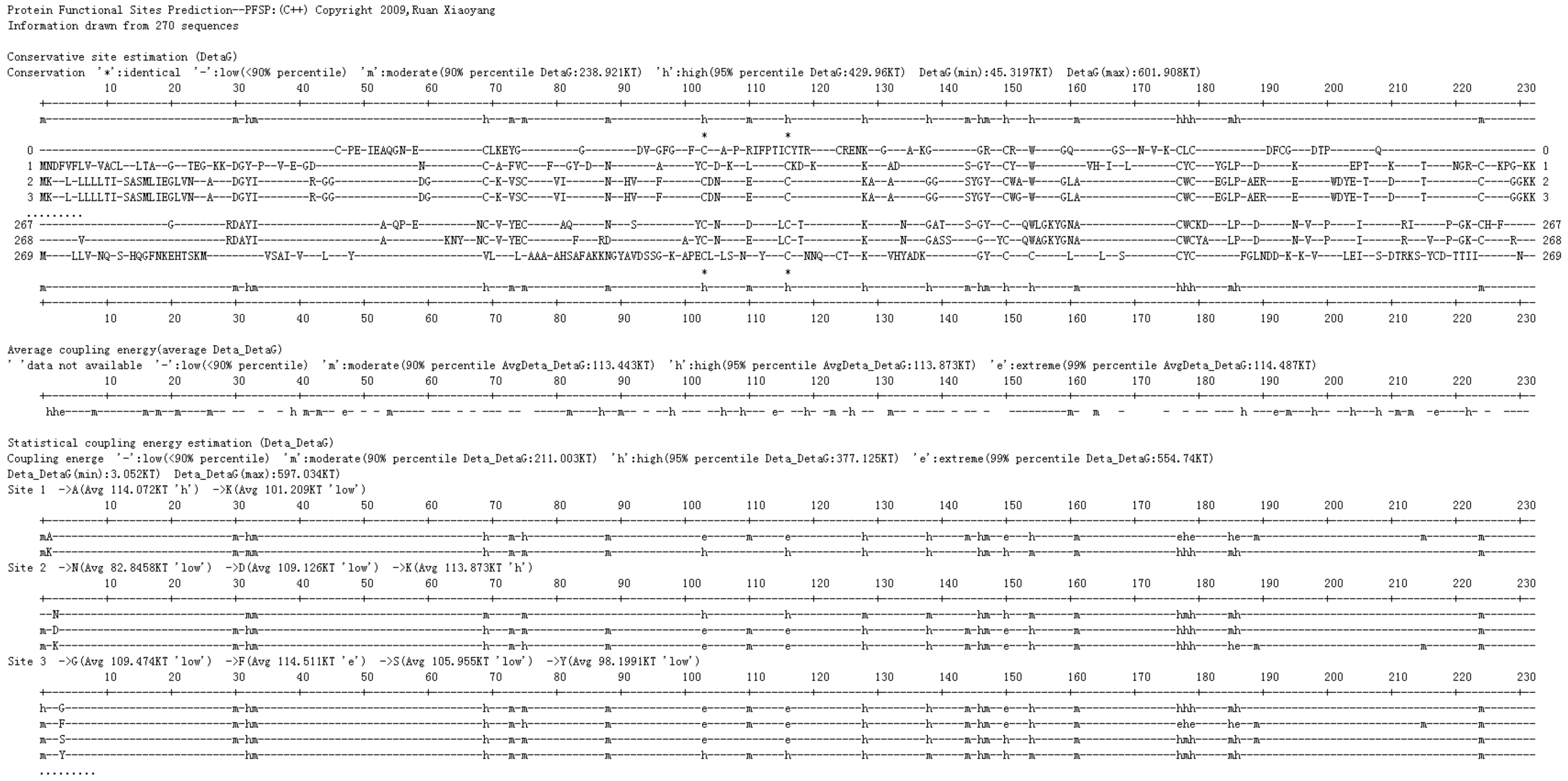
C:\PFSP_270_10 means [PFSP] _ [270 sequences] _ [sequence number threshold]

Sequence number threshold defined the minimum difference in sequence number before and after mutation (too little difference results in overestimation/underestimation of the actually effect). Generally, higher threshold means better statistics but may also miss sites with rare mutation. We recommend 10 as default value. However, if more sequences (such as 1000) are available, a higher threshold could be used.

From figure 18 (or sample result in ***SAMPLE_PFSP_270_10***), we can identify sites with different conservative levels as well as an estimated average coupling energy change for mutations at each site. These results enable you to estimate the global effect of mutation at specific site on the protein sequence. According to result ***SAMPLE_PFSP_270_10,*** mutations at site 3 to F (phenyl alanine), 47 to I (isoleucine), 114 to S (serine), 192 to Y (tyrosine) and 217 to C (cysteine) have extreme (>99th percentile) change in average coupling energy. These sites are potential candidates for your mutation experiment plan. However, this does not mean you can sit back and relax, for these ratiocinations are drawn from limited number of homologous protein sequences and only have statistical significance. On the other hand, differing parameters at each step may result in different results. For instance, by choosing Bayesian exact distance (BLAST local distance), we will identify mutations at site 31 to V (valine), 52 to S (serine), 56 to I (isoleucine), 72 to K (lysine), 223 to Q (glutamine) that have extreme values. This at first glance is very different from the result by applying approximate (Poisson) rule. After careful comparison, we can find some consistence between the two results. Bayesian exact rule identified the following mutations with significantly high average coupling energy change: site 3 to F (>95th), 56 to I (>99th), 123 to S (>95th). These sites actually have same positioning with site 3, 47, 114 respectively in the Poisson method (site ID difference was caused by different gap length introduced in different methods). Similarly, mutation at site 208 to Q has level >95th in the Poisson method, which corresponds to the mutation at 223 to Q in the Bayesian method. However, several sites failed to find corresponding records in the cross-validation. This is mostly caused by alignment error or insufficient number of sequences.

When limited number of sequence is available, it is always a good idea to do cross-validation between two or more methods. This can potentially avoid significance arise from error. Also, DO NOT merely pay attention to sites with extreme (>99th percentile) value, there isn't much difference between >95th and >99th considering the errors introduced at the homologous sequence selection step and the following parameter selection steps.

Figure 18



```
Protein Functional Sites Prediction--PFSP:(C++) Copyright 2009,Ruan Xiaoyang
Information drawn from 270 sequences

Conservative site estimation (DetaG)
Conservation  '*':identical  '-':low(<90% percentile)  'm':moderate(90% percentile DetaG:238.921KT)  'h':high(95% percentile DetaG:429.96KT)  DetaG(min):45.3197KT)  DetaG(max):601.908KT)
               10        20        30        40        50        60        70        80        90        100       110       120       130       140       150       160       170       180       190       200       210       220       230
     +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--
     m-----------------------------m-hm-----------------------------------h---m-m------------m--------------h------m-----h-----------h---------h-----m-hm--h---h------m---------------hhh-----mh-------------------------------------m--------
                                                                                                                   *            *
   0 ----------------------------------------------C-PE-IEAQGN-E----------CLKEYG---------G--------DV-GFG--F-C--A-P-RIFPTICYTR----CRENK--G---A-KG-------GR--CR--W----GQ------GS--N-V-K-CLC-----------DFCG---DTP-------Q------------------------ 0
   1 MNDFVFLV-VACL--LTA--G--TEG-KK-DGY-P--V-E-GD----------------N---------C-A-FVC---F--GY-D--N-------A-----YC-D-K--L-----CKD-K-------K---AD----------S-GY--CY--W--------VH-I--L-------CYC---YGLP--D-----K--------EPT---K----T----NGR-C--KPG-KK 1
   2 MK--L-LLLLTI-SASMLIEGLVN--A---DGYI--------R-GG-------------DG--------C-K-VSC----VI------N--HV---F------CDN----E-----C-----------KA--A-----GG----SYGY--CWA-W----GLA---------------CWC---EGLP-AER----E-----WDYE-T---D----T--------C----GGKK 2
   3 MK--L-LLLLTI-SASMLIEGLVN--A---DGYI--------R-GG-------------DG--------C-K-VSC----VI------N--HV---F------CDN----E-----C-----------KA--A-----GG----SYGY--CWG-W----GLA---------------CWC---EGLP-AER----E-----WDYE-T---D----T--------C----GGKK 3
   .........
 267 --------------------G--------RDAYI-------------------A-QP-E---------NC-V-YEC----AQ------N---S---------YC-N----D----LC-T---------K-----N---GAT---S-GY--C--QWLGKYGNA---------------CWCKD---LP--D-----N-V--P----I------RI-----P-GK-CH-F----- 267
 268 ------V----------------------RDAYI-------------------A---------KNY--NC-V-YEC-------F---RD-----------A-YC-N----E----LC-T---------K-----N---GASS----G--YC--QWAGKYGNA---------------CWCYA---LP--D-----N-V--P----I------R---V--P-GK-C----R--- 268
 269 M----LLV-NQ-S-HQGFNKEHTSKM---------VSAI-V---L---Y--------------------VL---L-AAA-AHSAFAKKNGYAVDSSG-K-APECL-LS-N--Y---C--NNQ--CT--K---VHYADK--------GY--C---C-----L----L--S--------CYC-------FGLNDD-K-K-V----LEI--S-DTRKS-YCD-TTII------N-- 269
                                                                                                                   *            *
     m-----------------------------m-hm-----------------------------------h---m-m------------m--------------h------m-----h-----------h---------h-----m-hm--h---h------m---------------hhh-----mh-------------------------------------m--------
     +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--
               10        20        30        40        50        60        70        80        90        100       110       120       130       140       150       160       170       180       190       200       210       220       230

Average coupling energy(average Deta_DetaG)
' 'data not available  '-':low(<90% percentile)  'm':moderate(90% percentile AvgDeta_DetaG:113.443KT)  'h':high(95% percentile AvgDeta_DetaG:113.873KT)  'e':extreme(99% percentile AvgDeta_DetaG:114.487KT)
               10        20        30        40        50        60        70        80        90        100       110       120       130       140       150       160       170       180       190       200       210       220       230
     +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--
      hhe----m-------m-m--m----m-- --  -  - h m-m-- e- - - m----- --- - - ---  --  -----m----h--m-- - --h --- --h--h--- e- --h- -m -h --  m-- - --- - -- -   ---------m-  m   -      -  - -- --- h ---e-m---h-- --h---h -m-m  -e---h-  -  ----

Statistical coupling energy estimation (Deta_DetaG)
Coupling energe  '-':low(<90% percentile)  'm':moderate(90% percentile Deta_DetaG:211.003KT)  'h':high(95% percentile Deta_DetaG:377.125KT)  'e':extreme(99% percentile Deta_DetaG:554.74KT)
Deta_DetaG(min):3.052KT)  Deta_DetaG(max):597.034KT)
Site 1  ->A(Avg 114.072KT 'h')  ->K(Avg 101.209KT 'low')
               10        20        30        40        50        60        70        80        90        100       110       120       130       140       150       160       170       180       190       200       210       220       230
     +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--
     mA----------------------------m-hm-----------------------------------h---m-h------------m--------------e------m-----e-----------h---------h-----m-hm--e---h------m---------------ehe-----he--m-------------------------m--------m--------
     mK----------------------------m-mm-----------------------------------h---m-m------------m--------------h------------h-----------h---------h-------hm--h---m------m---------------hhh-----mh-------------------------------------m--------
Site 2  ->N(Avg 82.8458KT 'low')  ->D(Avg 109.126KT 'low')  ->K(Avg 113.873KT 'h')
               10        20        30        40        50        60        70        80        90        100       110       120       130       140       150       160       170       180       190       200       210       220       230
     +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--
     --N-----------------------------mm-----------------------------------m-----m---------------------------h------------h-----------m---------m-------hm--h---m------m---------------hmh-----mh-------------------------------------m--------
     m-D---------------------------m-hm-----------------------------------h---m-m------------m--------------e------m-----e-----------h---------h-----m-hm--e---h------m---------------hhh-----mh-------------------------------------m--------
     m-K---------------------------m-hm-----------------------------------h---m-h------------m--------------e------m-----e-----------h---------h-----m-hm--e---h------m---------------hhh-----he--m-------------------------m--------m--------
Site 3  ->G(Avg 109.474KT 'low')  ->F(Avg 114.511KT 'e')  ->S(Avg 105.955KT 'low')  ->Y(Avg 98.1991KT 'low')
               10        20        30        40        50        60        70        80        90        100       110       120       130       140       150       160       170       180       190       200       210       220       230
     +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--
     h--G--------------------------m-hm-----------------------------------h---m-m------------m--------------e------m-----e-----------h---------h-----m-hm--e---h------m---------------hhh-----mh-------------------------------------m--------
     m--F--------------------------m-hm-----------------------------------h---m-h------------m--------------e------m-----e-----------h---------h-----m-hm--e---h------m---------------ehe-----he--m-------------------------m--------m--------
     m--S--------------------------m-hm-----------------------------------h---m-m------------m--------------e------------e-----------h---------h-----m-hm--h---h------m---------------hmh-----mh--m----------------------------------m--------
     m--Y----------------------------hm-----------------------------------h---m-m------------m--------------h------m-----h-----------h---------m-----m-hm--h---m------m---------------hmh-----mh-------------------------------------m--------
     .........
```

Protein functional sites prediction results. It is composed of three parts. Part 1 (Conservative site estimation) showed three levels of conservation labeled with '-', 'm', 'h' (for low, medium, high conservation). Part 2 (Average coupling energy) showed the averaged coupling energy changes at the remaining sites when each site mutated. Blank area means no mutation data is available for that site. The other labels represented the highest average coupling energy seen in that site. Part3 (Statistical coupling energy) presented the mutation data, if any, for each site. It also outputted the average coupling energy for each possible mutation target AA.

Genetic Distance

To have a better understanding of this section, the following knowledge might be helpful.

PAM distance

There are two ways to calculate genetic distance -> approximate method and exact method. **Approximate method** gives a quick estimation on the evolutionary distance by simply calculating the proportion of difference AAs between two aligned protein sequences. Additional correction algorithm was used to minimize the gap between approximate and exact method.

Simple p-distance

As noted above, the p-distance is merely the proportion of different amino acids between two sequences compared. It is used without additional correction step.

$$p = n_d/n$$

$$V(p) = p(1 - p)/n$$

Here n_d and n are the number of amino acid differences and the total number of amino acids compared, respectively.

Poisson-correction distance

This distance is for estimating the number of amino acid substitutions per site under the assumption that the number of amino acid substitutions at each site follows the Poisson distribution. This estimator (a) and its variance are given by

$$d = -\log_e(1 - p)$$

$$V(d) = p/[(1 - p)n]$$

where *p* is estimated by equation given in simple p-distance.

Gamma distance

This distance is an estimate of the number of amino acid substitutions per site under the assumption that the rate of amino acid substitution varies from site to site and follows the gamma distribution with parameter *a*. This distance and its variance can easily be computed from Nei *et al.*'s (1995) [6] work.

$$d = a[(1 - p)^{-1/a} - 1]$$

$$V(d) = p[(1 - p)^{-(1 + 2/a)}]/n$$

When *a* = 2 is used, *d* is close to Dayhoff's (1978) [1] PAM distance per site (0.01 PAM)

Bayesian exact distance calculates the probability score distribution from PAM 1 to PAM 400 (or higher) [7]. Then a relative probability was calculated to make all probability scores sum to 1. The evolutionary distance (in PAM/100) was obtained by computing the expectation value *E(X)*

$$\boldsymbol{d = E(X) = \sum_{pam=1}^{400} S_{pam} P_{pam}}$$

$$\boldsymbol{V(d) = E(X^2) - E^2(X)}$$

Where $\boldsymbol{S_{pam}}$ is the probability score at distance ***pam*** and $\boldsymbol{P_{pam}}$ is the relative probability. Figure 19 illustrated the relative p distribution from PAM 1 to PAM 270 (Bayesian, BLAST local rule) by aligning the following 2 sequences, whereas Figure 20 used Bayesian Needleman global rule.

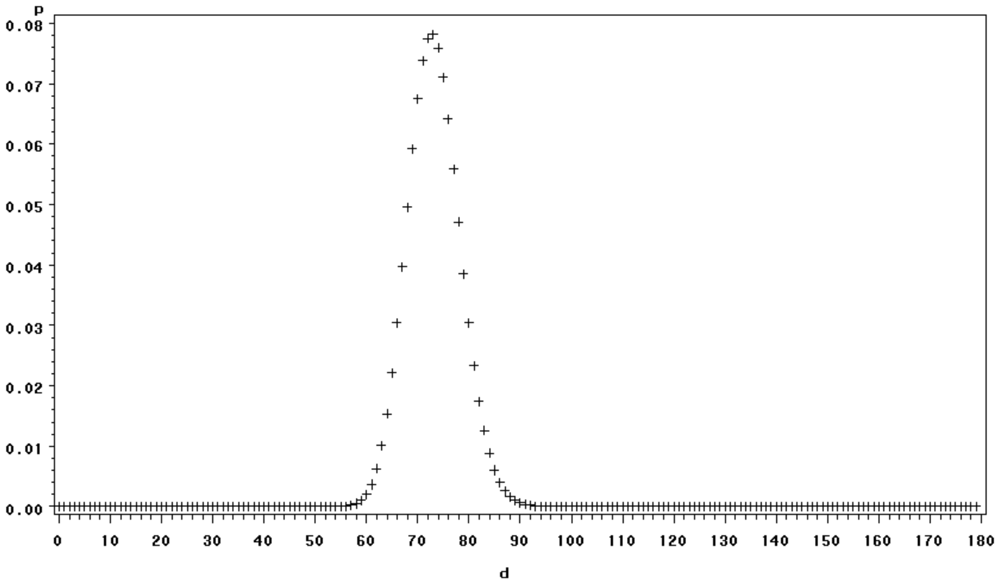
>gi|159162458|pdb|1I6F|Insect-Specific Neurotoxin Variant 5 (Cse-V5)

KDGYPVDSKGCKLSCVANNYCDNQCKMKKASGGHCYAMSCYCEGLPENAKVSDSATNICG

>gi|158931147|sp|P60213.2|SC49A_TITCA Toxin Tc48b/Tc49a

KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTWSRSTNRCGR

Figure 19



In table 3, Gamma algorithm with a=2 gives very similar result as Bayesian exact (Needleman global). We noticed the highest distance when using BLAST local rule. This is because the rest of algorithms maximize the alignment score on a global range, whereas BLAST local rule only consider segments with extension scores above significance threshold. However, different distance estimation methods will give same results if two sequences are very similar.

Figure 20

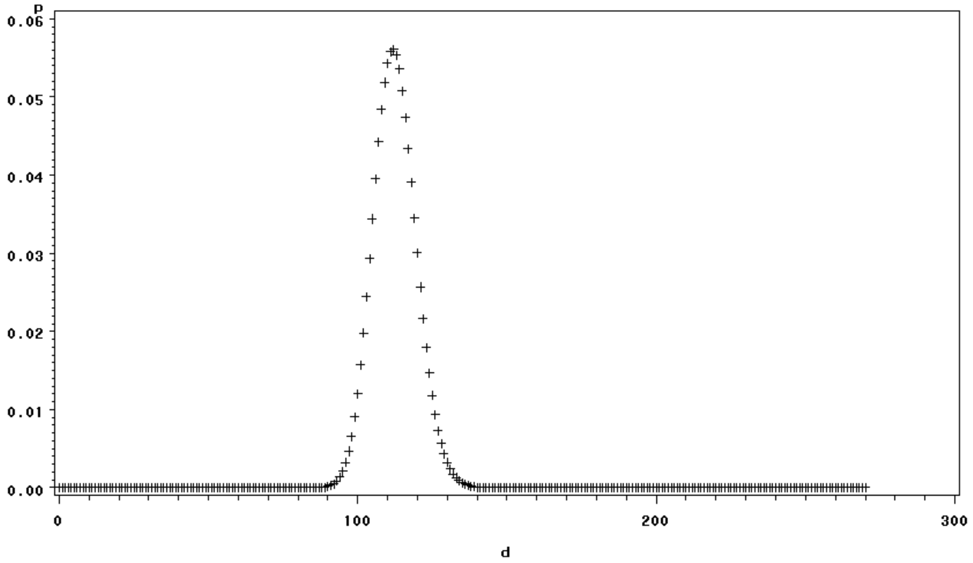


Table 3

Distance estimation type	Distance (PAM)	Variance (PAM)
Simple P-distance	48	0.201
Poisson correlation distance	66	0.756
Gamma distance (a=2)	78	1.465
Bayesian exact (Needleman global)	78	2.917
Bayesian exact (BLAST local)	119	5.878

The ideal way to calculate Bayesian exact distance is try the scoring matrices one by one (from PAM 1 to 400 or higher) on the alignment sequences and identify the one maximizes the alignment score. However, this approach requires prohibitively long time and is not practical when pair wise distance need to be calculated for several hundreds of sequences. In PFSP, when Bayesian exact algorithm was selected, Gaston scoring matrix [8] was used to guide the initial alignment. Then PAM matrices were then tried on the alignment result to identify the one maximized the score. Gaston scoring matrix was derived from an exhaustive matching of the entire MIPS (Munich Information Center for Protein Sequences) database and was recommended as the initial scoring matrix before subsequent refinement was applied.

Bayesian step controlled how PAM matrices were applied to the alignment result. A Bayesian step with value *N* means PAM 1, 1+*N*, 1+2*N*… will be tried in order. Generally, *N*=10 have very similar distance and variance result with *N*=1.

Phylogenetic Tree

A phylogenetic tree or evolutionary tree is a tree showing the evolutionary relationships among various biological species or other entities that are believed to have a common ancestor. In a phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants, and the edge lengths in some trees correspond to time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units (HTUs) as they cannot be directly observed. Generally, there are two types of Phylogenetic tree: rooted and un-rooted. A rooted phylogenetic tree is a directed tree with a unique node represents the most recent common ancestor of all the terminals of the tree. Un-rooted trees do not need to make assumption about a common ancestry. Both rooted and unrooted phylogenetic trees can be either bifurcating or multifurcating. A bifurcating tree has a maximum of two descendents arising from each interior node, while a multifurcating tree may have more than two. Current version of PFSP calculates un-rooted tree from pair wise distance matrix by using the neighbor-joining (NJ) method developed by Saitou,N et al. [9] and later modified by Studier,J.A. et al [10]. The result was then used to guide the multiple sequence alignment. An apparent advantage of NJ method lies in its ability to correctly reconstruct the tree from distance matrix D given that the distances in D correspond exactly to those in an actual tree. This property of NJ makes it the best choice to compute phylogenetic tree for large amount of homologous protein.

To construct NJ tree, the pair of sequences which minimizes the total distance between all nodes in the distance matrix was marked as neighbor. This is equal to find the pair of sequences that shares the longest path to the remaining nodes, which can be mathematically attained by minimizing the following term

$$S_{ij} = (N-2)D_{ij} - R_i - R_j$$

Where D_{ij} is the distance between node *i* and *j*. R_i is the sum of distance from node i to all remaining nodes. *N* is the number of nodes.

$$R_i = \sum_{k \neq i} D_{ik}$$

This pair of sequences was then merged into a new node *u*. The distance between *u* and other nodes can be calculated by

$$D_{iu} = \frac{1}{2}\left(D_{ik} + D_{jk} - D_{ij}\right) \text{ for } k \neq i, j$$

The branch lengths from *i* to *u* and *j* to *u* are

$$D_{iu} = \frac{1}{2(N-2)}\left[(N-2)D_{ij} + R_i - R_j\right]$$

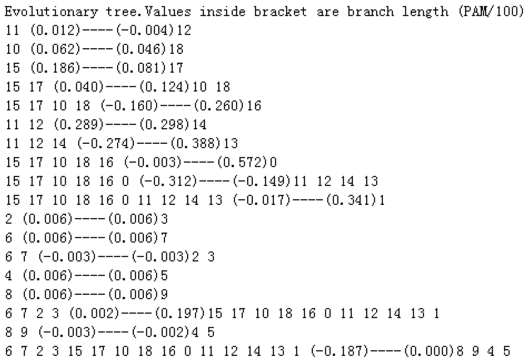
And

$$D_{ju} = \frac{1}{2(N-2)}[(N-2)D_{ij} + R_j - R_i]$$

NJ method has a time complexity of *O (n^2)* to find the neighbor and *O (n)* to update the R_i array. In PFSP, the NJ tree calculation module has been fine-tuned to handle several hundreds of sequences within few seconds.

The NJ result outputted by current version of PFSP has the following appearance

Figure 21



```
Evolutionary tree.Values inside bracket are branch length (PAM/100)
11 (0.012)----(-0.004)12
10 (0.062)----(0.046)18
15 (0.186)----(0.081)17
15 17 (0.040)----(0.124)10 18
15 17 10 18 (-0.160)----(0.260)16
11 12 (0.289)----(0.298)14
11 12 14 (-0.274)----(0.388)13
15 17 10 18 16 (-0.003)----(0.572)0
15 17 10 18 16 0 (-0.312)----(-0.149)11 12 14 13
15 17 10 18 16 0 11 12 14 13 (-0.017)----(0.341)1
2 (0.006)----(0.006)3
6 (0.006)----(0.006)7
6 7 (-0.003)----(-0.003)2 3
4 (0.006)----(0.006)5
8 (0.006)----(0.006)9
6 7 2 3 (0.002)----(0.197)15 17 10 18 16 0 11 12 14 13 1
8 9 (-0.003)----(-0.002)4 5
6 7 2 3 15 17 10 18 16 0 11 12 14 13 1 (-0.187)----(0.000)8 9 4 5
```

Sequences were represented by their corresponding ID in the original file. Negative lengths can be simply replaced with zero.

Construct Custom BLOSUM Scoring Matrix

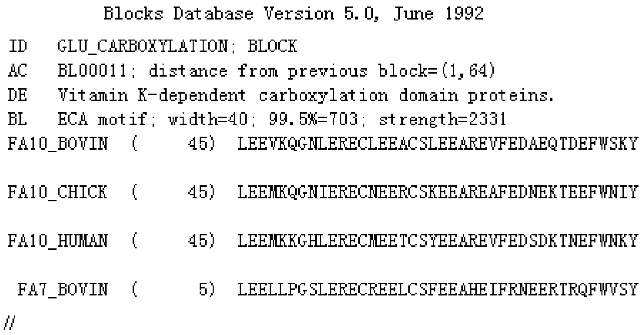
BLOSUM (BLOcks of amino acid Substitution Matrix) is a scoring matrix used for protein sequence alignment. Different from PAM scoring matrix, BLOSUM is based on local alignments. The idea of BLOSUM was put forward by Henikoff et al [2] in 1992. They scanned the BLOCKS database for very conserved regions of protein families (without gap) and then counted the relative frequencies of AAs and their substitution probabilities. Then, they calculated a log-odds for each of the 210 possible substitutions of the 20 standard AAs. All BLOSUM are based on observed alignments.

With one initial block database file, several sets of BLOSUM matrices can be derived according to different re-clustering levels. This level determined the similarity threshold above which two or more sequences will be merged into one single sequence and then comparing those sequences (that were all more divergent than the given level) only; thus reducing the contribution of closely related sequences. As such BLOSUM80 is used for less divergent alignments, whereas BLOSUM45 is used for more divergent one. This is just the opposite of PAM.

In PFSP, BLOSUM database (from BLOSUM100 to BLOSUM14) derived from standard block file (blocks-5.0.dat. Available at http://blocks.fhcrc.org/blocks/uploads/blosum/) has been pre-calculated and is ready for use. The original block file collected information from 2106 blocks and is competent for alignment work under most circumstances. However, users are allowed to construct their own BLOSUM matrix and use the custom scoring matrix in alignment.

A block file should have the following format.

Figure 22



```
          Blocks Database Version 5.0, June 1992
 ID    GLU_CARBOXYLATION; BLOCK
 AC    BL00011; distance from previous block=(1,64)
 DE    Vitamin K-dependent carboxylation domain proteins.
 BL    ECA motif; width=40; 99.5%=703; strength=2331
 FA10_BOVIN  (      45)  LEEVKQGNLERECLEEACSLEEAREVFEDAEQTDEFWSKY

 FA10_CHICK  (      45)  LEEMKQGNIERECNEERCSKEEAREAFEDNEKTEEFWNIY

 FA10_HUMAN  (      45)  LEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKY

  FA7_BOVIN  (       5)  LEELLPGSLERECREELCSFEEAHEIFRNEERTRQFWVSY
//
```

File format of block. New block file can be downloaded from http://blocks.fhcrc.org/blocks/uploads/blosum/

There should be a title describes the version of the block file. The first word of the title must be "Blocks" (case sensitive). Such as

Blocks hello

Blocks 3000 block

Blocks custom version 1

are all allowed. But

blocks hello

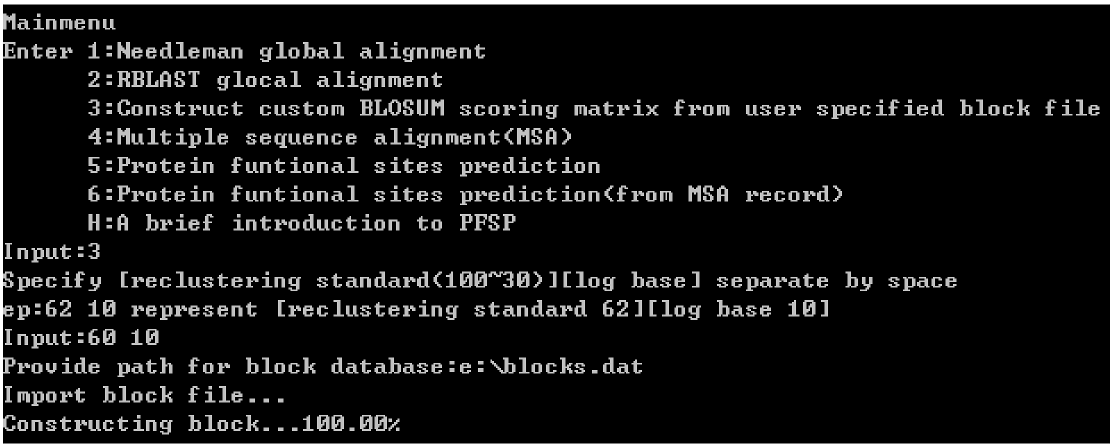
My block 3000

are not allowed. (*Note: PFSP will warn about the use of wrong title, you can choose to proceed if you insist to use wrong title*)

Each block should have ID, AC, DE, BL information in separated line. “//” was used to mark the end of a block. (*Note: Block file information is case and newline sensitive but blank insensitive*)

To create custom BLOSUM matrix, choose option 3, input re-clustering standard and log base value. Then provide the path to custom block file. Here is a sample output of custom BLOSUM60 matrix named with “Blocks hello my new BLOSUM”

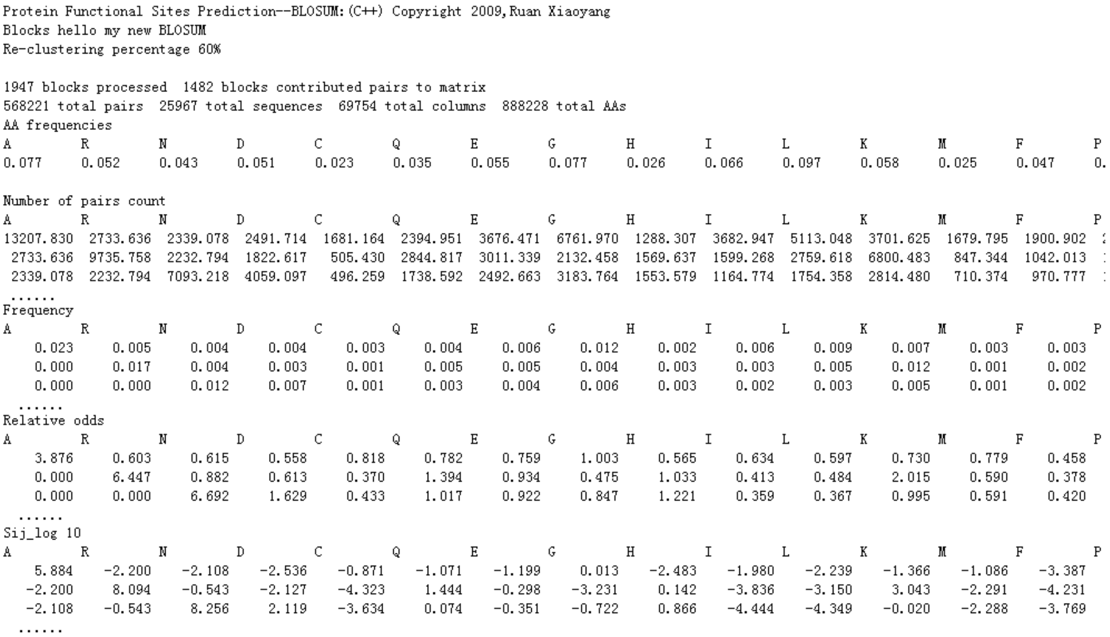
Figure 23



```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      H:A brief introduction to PFSP
Input:3
Specify [reclustering standard(100~30)][log base] separate by space
ep:62 10 represent [reclustering standard 62][log base 10]
Input:60 10
Provide path for block database:e:\blocks.dat
Import block file...
Constructing block...100.00%
```

The output of BLOSUM file was shown in figure 24

Figure 24



```
Protein Functional Sites Prediction--BLOSUM:(C++) Copyright 2009,Ruan Xiaoyang
Blocks hello my new BLOSUM
Re-clustering percentage 60%

1947 blocks processed  1482 blocks contributed pairs to matrix
568221 total pairs  25967 total sequences  69754 total columns  888228 total AAs
AA frequencies
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
0.077     0.052     0.043     0.051     0.023     0.035     0.055     0.077     0.026     0.066     0.097     0.058     0.025     0.047     0.

Number of pairs count
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
13207.830  2733.636  2339.078  2491.714  1681.164  2394.951  3676.471  6761.970  1288.307  3682.947  5113.048  3701.625  1679.795  1900.902
 2733.636  9735.758  2232.794  1822.617   505.430  2844.817  3011.339  2132.458  1569.637  1599.268  2759.618  6800.483   847.344  1042.013
 2339.078  2232.794  7093.218  4059.097   496.259  1738.592  2492.663  3183.764  1553.579  1164.774  1754.358  2814.480   710.374   970.777
 ......
Frequency
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
    0.023     0.005     0.004     0.004     0.003     0.004     0.006     0.012     0.002     0.006     0.009     0.007     0.003     0.003
    0.000     0.017     0.004     0.003     0.001     0.005     0.005     0.004     0.003     0.003     0.005     0.012     0.001     0.002
    0.000     0.000     0.012     0.007     0.001     0.003     0.004     0.006     0.003     0.002     0.003     0.005     0.001     0.002
  ......
Relative odds
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
    3.876     0.603     0.615     0.558     0.818     0.782     0.759     1.003     0.565     0.634     0.597     0.730     0.779     0.458
    0.000     6.447     0.882     0.613     0.370     1.394     0.934     0.475     1.033     0.413     0.484     2.015     0.590     0.378
    0.000     0.000     6.692     1.629     0.433     1.017     0.922     0.847     1.221     0.359     0.367     0.995     0.591     0.420
  ......
Sij_log 10
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
    5.884    -2.200    -2.108    -2.536    -0.871    -1.071    -1.199     0.013    -2.483    -1.980    -2.239    -1.366    -1.086    -3.387
   -2.200     8.094    -0.543    -2.127    -4.323     1.444    -0.298    -3.231     0.142    -3.836    -3.150     3.043    -2.291    -4.231
   -2.108    -0.543     8.256     2.119    -3.634     0.074    -0.351    -0.722     0.866    -4.444    -4.349    -0.020    -2.288    -3.769
  ......
```

To use custom BLOSUM matrix in alignment, just replace the original block file ***Blocks.dat*** in C:\\WINDOWS with your own block file (rename to **Blocks.dat**). You can also restore the original BLOSUM matrix by replacing the custom block file with the original one.

Acknowledgement

I would like to express my special gratitude to Mrs Zhu Jun who gave me the ideas and materials. Her help makes it possible for me to start the work on PFSP.

References

1. **Dayhoff MO, R.V.Eck, C.M.Park. A model of evolutionary change in proteins. Atlas of protein sequence and structure 1972; 5:89-99.**

2. **Henikoff S, Henikoff,JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992; 89:10915-9.**

3. **Needleman SB, Wunsch,CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48:443-53.**

4. **Altschul SF, Gish,W, Miller,W, Myers,EW, Lipman,DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10.**

5. **Lockless SW, Ranganathan,R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999; 286:295-9.**

6. **Grishin NV. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 1995; 41:675-9.**

7. **Agarwal P, States,DJ. A Bayesian evolutionary distance for parametrically aligned sequences. J Comput Biol 1996; 3:1-17.**

8. **Gonnet GH, Cohen,MA, Benner,SA. Exhaustive matching of the entire protein sequence database. Science 1992; 256:1443-5.**

9. **Saitou N, Nei,M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4:406-25.**

10. **Studier JA, Keppler,KJ. A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol 1988; 5:729-31.**